ihmc

W911NF-11-2-0015

Subject: Transmittal Letter for Final Report

Title: **Extensible Collaborative Systems for Mission Planning**

IHMC Principal Investigator: Dr. James Allen

Submitted to: Dr. Timothy Hanratty
Army Research Lab
ATTN: RDRL-CII-C
Aberdeen Proving Ground
Aberdeen, MD, 21005

Dear Dr. Hanratty,

The Florida Institute for Human and Machine Cognition (IHMC) — a not-for-profit research institute of the state university system of Florida — is pleased to submit the final report for the project **Extensible Collaborative Systems for Mission Planning.**

If you have any questions, please do not hesitate to contact Dr. Allen on technical issues at jallen@ihmc.us. For administrative or contracting questions, please contact me at 850-202-4473 or dthacker@ihmc.us.

Thank you for your consideration.

Diana Thacker

Diana Thacker
Director for Grants and Contracts

Final Technical Report

W911NF-11-2-0015

Submitted by: Dr. James Allen

1. Goals

The goal of this work was to develop an Extensible Agent for Collaborative Tasks (EXACT). EXACT was to embody a novel computational theory of collaboration that drives the interactions between humans and machines, as they work collaboratively to construct and manage mission plans. EXACT was to enable a multimodal, dialogue-based interaction that is more intuitive and effective for both novice and expert users alike. EXACT would make mission planning systems act more like good human assistants than passive pieces of software.

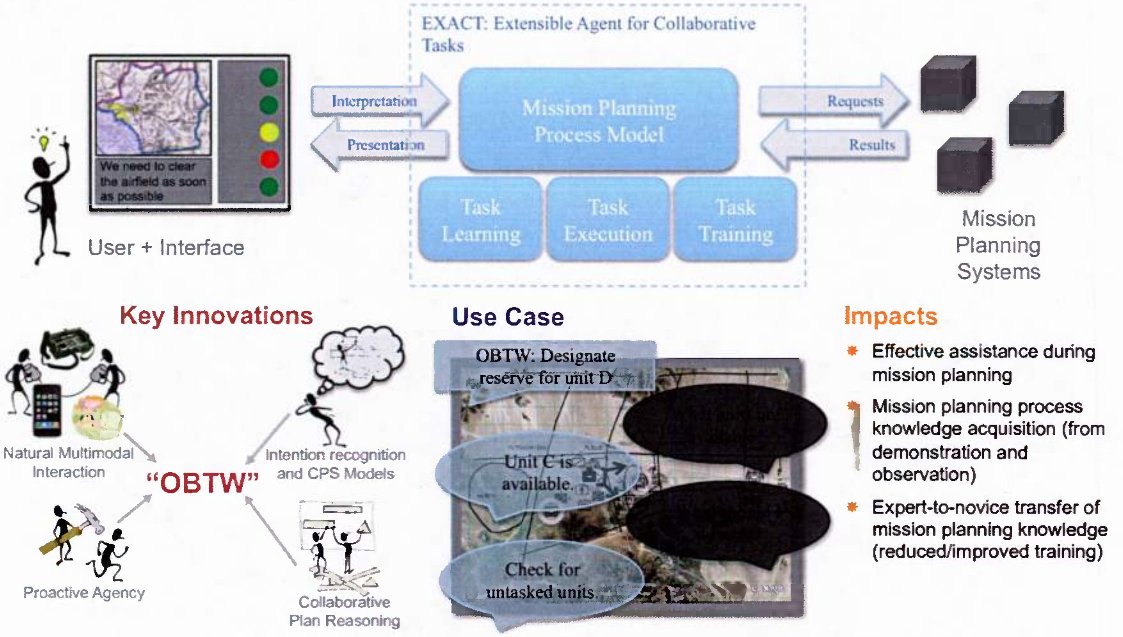
EXACT: Extensible Agent for Collaborative Tasks
Mission Planning Process Model
Interpretation
Presentation
Requests
Results
Task Learning
Task Execution
Task Training
We need to clear the airfield as soon as possible
User + Interface
Mission Planning Systems
Key Innovations
Natural Multimodal Interaction
Intention recognition and CPS Models
"OBTW"
Proactive Agency
Collaborative Plan Reasoning
Use Case
OBTW: Designate reserve for unit D
Unit C is available.
Check for untasked units.
Impacts
Effective assistance during mission planning
Mission planning process knowledge acquisition (from demonstration and observation)
Expert-to-novice transfer of mission planning knowledge (reduced/improved training)

Figure 1: The ExACT Concept

Given that the OBTW program was cancelled at the end of its first year, these goals were not met. We did, however, build a significant infrastructure for future research in this area and performed a pilot experiment to explore possible evaluation measures.

2. The System

We built a first prototype of the ExACT system as shown in Figure 2. While several of the key components were adapted from our previous work, several significant components were built

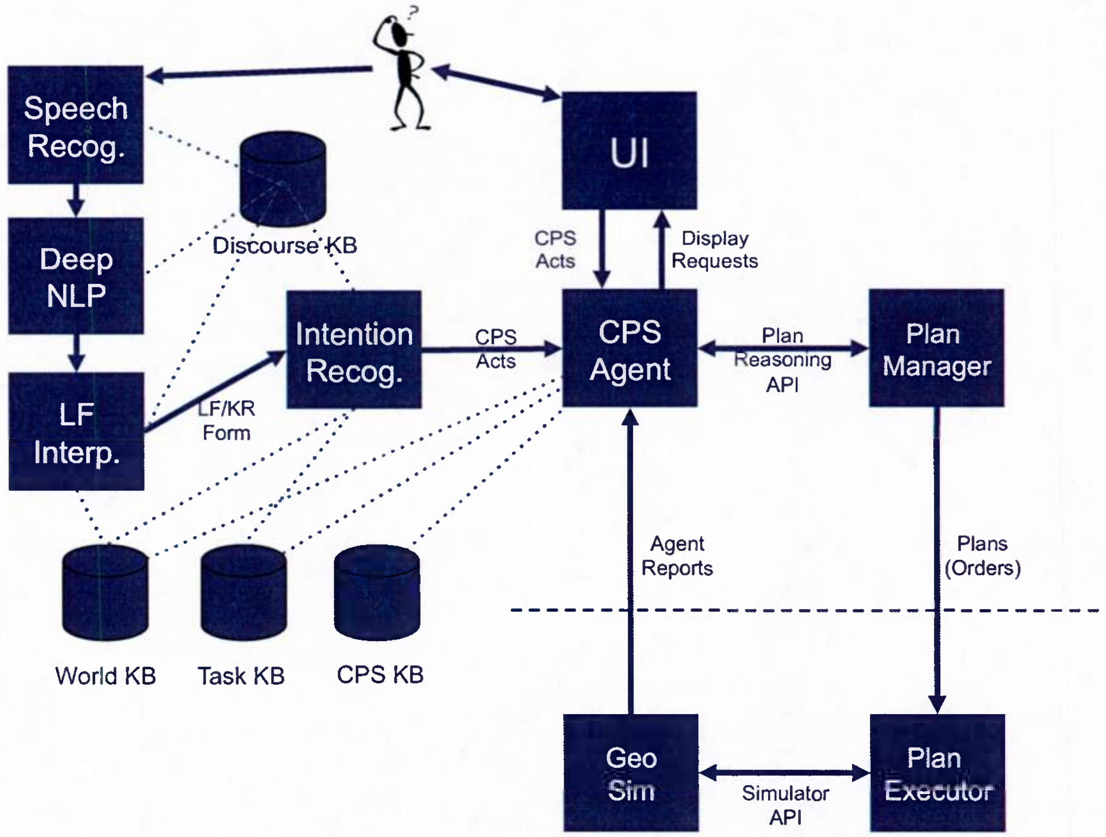

Figure 2: The ExACT System

from scratch expressly for the project. The first of these was the ExACT GUI, which the user can use to interactively build plans and manage their execution. The GUI was defined to allow the full range of planning operations that the system could perform. In addition, every GUI action could also be specified using the spoken natural language interface. This architecture was designed to allow for careful evaluation experiments where we could vary the available modalities and level of helpful behavior by the system. The second major component was GeoSim, a world simulator that modeled an unfolding natural disaster and simulated the actions of agents that were following plans built using the system.

3. Experimental setup

At the cnd of the first ycar, we performed some preliminary experiments to explore various ways to cvaluate the benefit of using Oh-By-The-Way behaviors.

We used a Wizard-of-Oz setup, where:

- the system was run off the wizard's laptop;

- the user interface was displayed on a large screen external display (47"); significantly, the GUI was designed to show limited information about the world, to encourage users to ask questions;

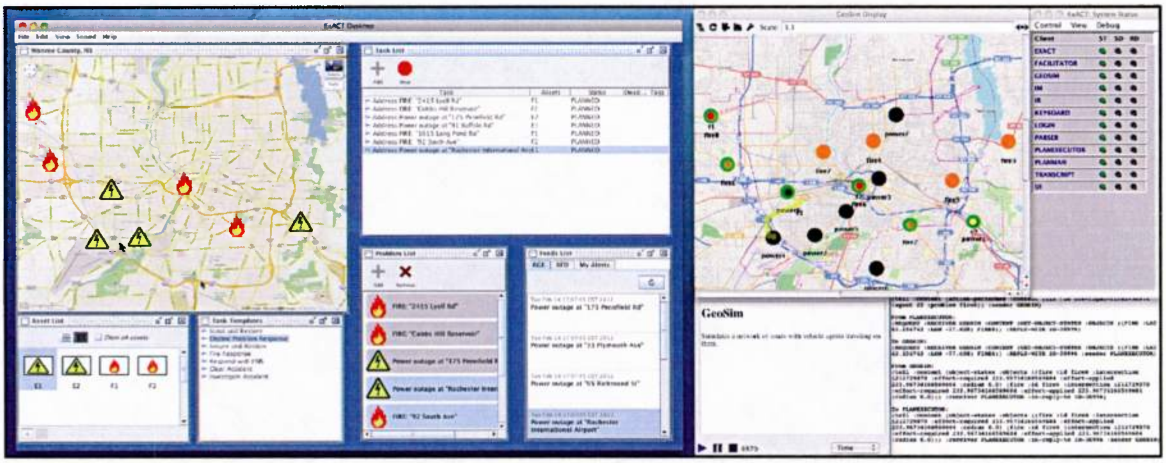

Figure 3. Uscr display (left, blue background) and wizard display (right).

- the wizard, in addition to seeing and controlling the user display, also had a private display providing access to information about the world (from the simulator) not shown on the user's display.

The user would only interact with the system via spokcn language. A lapel wireless microphone was used to collect the user's speech. The wizard used a mouse to affect changes to the user interface according to the user's instructions. The wizard would also answer users' questions and, on occasion, proactively offer suggestions, via spoken language. The wizard's speech was also collected, via a headset microphone, into a separatc audio channel from the user's speech. A third audio channel recorded system outputs (alerts).

We recorded a video of both thc user's display and the wizard's display (shown side-by-side). Figure 1 shows a still from the video, with the user display on the left (blue background), and the wizard display on the right.

Subjects

For this phase of the experiment, given the time constraints and the cxploratory nature of the experiment, we only used four volunteer subjects. There were two female and two male subjects; none were familiar with our systcm, and none had acted before as dispatchers in emergency operations or performed tasks similar in nature.

A few preliminary tests were run prior to the actual expcriment, to tune the parameters of the experiment (duration of experiment, number of events, speed-up factor); the four subjects had not participated in these preliminary tests.

Experimental protocol

A session with a subject was divided into three distinct phases. In each phase, a different task was loaded; each task included 8 fires and 8 power outages. Of note, we designed the tasks so that all problems of the same type required equal amounts of cffort and resolving fires required 4 times as much effort as fixing electric problems. The subject was to try and resolve as many problems as possible within a certain amount of time.

The three phases were:

1. ***Dry run***: This phase started with a short tutorial in which the wizard explained the system's functionality, the user interface, and the task that the user was going to solve. The wizard demonstrated how to build and execute a plan for extinguishing a fire and fixing an electrical problem. Thereafter, the subject was encouraged to continue solving the remaining problems, as well as ask any questions they might have about the capabilities/functionality of the system. The wizard provided explanations/clarifications as appropriate. This portion of the phase, where the subject solved problems, was timed; this was to familiarize the uscr with the timer, but the subjects were advised to explore how the system worked rather than try to solve as many as possible in thc time availablc.

2. ***Test 1***: This was a 10-minute problem solving session with the "unhelpful" system. For this condition, the wizard always performed the user's command, and only answered questions that were about features of any particular object (goal/problem, task, crew). The user was asked to solve as many problcms as possible in the allotted time.

3. ***Test 2***: This was also a 10-minute problem solving session with the system. During this phase, two of the users (chosen randomly) interacted with an "unhelpful" system (as in Phase II), while the other two users interacted with a "helpful" system, in which thc wizard/system volunteered situation information that could bc helpful in obtaining a better solution (eg, whether crews are available or not, whether one is closer than another to the problem being solved; status of problems and crews, etc.).

Hypothesis

Our hypothesis was that the information provided by the helpful system would lead to better solutions. We further hypothesizcd that that would be reflected in a faster average time for finding a solution. Thcrcfore, as our primary measure of performance, we decided to use the average time to devclop a plan for solving a problem and dispatch a crew to execute the plan (in short, *average planning time*):

$$apt = T / N$$

where T is the time between thc start of thc experiment until the time of the last dispatch before the end of the 10-minute period; and N is the number of times crews were successfully dispatched to solve problems. Note that even though problems may not be solved until after the end of the 10-minute period, if the crews were dispatched before the time was up, we counted them as successful. Conversely, we did not give credit to any work pcrformed between the last successful dispatch of a crew and the end of the time allocated for the task.

Results and conclusion

The results for the 4 participants in the study are in Table 1.

	S1	S2	S3	S4
Test 1	00:52.329	00:36.475	00:33.292	00:44.881
Test 2	00:56.378	00:34.554	00:32.702	00:33.869
Change	*7.74%*	*-5.27%*	*-1.77%*	*-24.54%*

Table 1. Average planning times. For Test 2, S1 and S2 interacted with the "helpful" system, and S3 and S4 interacted with the "unhclpful" system. During Test 1, all four subject interacted with an unhelpful" system.

At first sight it may appear that the "helpful" system has worse performance than the "unhelpful" one. In fact, this is misleading, and reveals some important problems with our experimcntal setup and the performance measure we used.

First, it turns out that, because the system was sped up significantly to allow completion of a significant number of the 16 problems in the 10-minute period, the time to ask questions and/or consider options was disproportionately large compared to the time to send a crew across town on a very inefficient route. Several subjects quickly realized that speed trumped smarts in this game. Unfortunately, the subjects we had in the preliminary tests happened to be slower than the final test subjects, so we didn't catch this problem in time. Unfortunately, we were limited in the amount of time we could ask our volunteer subjects to spend with the system, but it is likely that we could have tuned the system to a more realistic state by slowing the simulator further and having fewer problems in each scenario. In reality, of course, the time to travel to location and solve the actual problem would exceed by far the time to plan the best course of action.

A second issue that we noticed was that the training done during the dry run (first phasc) was insufficient for novices to develop a strategy. Two of our subjects (S1, and S4) changed significantly their problem solving strategy between the first and the second test.

Finally, and as a consequence of the above issues, the pcrformance measure we chose was not ablc to capture the difference in magnitude between, on one hand, the time users allocated to situation assessment and to plan optimization -- the two aspects most affected in the "helpful" mode -- and, on the other hand, the time allocated to just issuing commands to the system to communicate their intent. As such, the average planning time turns out to be less than ideal as a performance mcasure for detecting the cffcct of the system's suggcstions on thc outcome of the users' problcm solving behavior.

Notwithstanding the above limitations of our initial experimental approach, we did obtain some interesting insights from this experimcnt. It turned out that most users were somewhat frustrated by the interaction with the system during the first test. At the end of it, several asked if the system could be more helpful, or if they could ask certain questions to find out more about the situation. During the interaction with the unhelpful system, users generally built plans for solving the emergency problems with little regard to how "good" these plans were. However, when the "helpful" system started making suggestions (e.g., better alternatives, or alerting the user to when resources became available -- see Table 2 for some examples), users immediately shifted their strategy towards trying to optimize the routes. Therefore, while this is not captured by the *apt* measure, and we didn't have a ready way to measure this, it appears that overall the routes are, indeed, significantly more efficient for both S1 and S2's solutions in Test 2 compared to their own solutions in Test 1 and those of S3 and S4 for Test 2. We are currently exploring the possibility of developing an additional performance measure that accounts for just the quality of the planned routes, and not the time spent on planning.

Utterance type	Goal
<crew> is closer to <problem>	to suggest a more efficient plan
<crew> is available / has finished their job	to make user aware of resource availability
<crew> is busy / still working on <problem>	to prevent user from over-allocating a resource
<problem> has already been resolved	to prevent user from sending a second crew

Table 2. Examples of types of unsolicited suggestions made by the "helpful" system.

At the end if Test 2, we asked S1 and S2 to compare their experience with the two modes of the system. They both considered the helpful system much easier to work with. In the words of one subject, "it was [...] and interaction as opposed to me shouting commands". Also, he appreciated "getting hints that I wouldn't have asked otherwise".

In conclusion, even though this exploratory experiment was carried out at such a small scale, we think it provided useful insights into how users react to helpful system behaviors. First, it appears that the users' problem solving strategy can be dramatically influenced by the type of information provided by the system. Second, it appears that a more helpful system will nudge users into finding objectively better solutions. And finally, users report much better satisfaction from interacting with a helpful system compared to an unhelpful one.

4. Deliverables and Tech Transfer

None - program was cancelled before any deliverables were scheduled.